DATA MINING AND MACHINE LEARNING PROJECT REPORT

Ameya Satish Dalal
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x18184430@student.ncirl.ie

Abstract **-Nowadays, a large amount of data is available everywhere on the internet. The most important thing is finding an insight by analysing the elements and getting the required information which will help us in decision made by considering data and enhancing the process with its workflow can be observed by using machine learning and data mining methods. Artificial Intelligence helps to provide potential support to the system with machine learning which is an intrinsic part without any confusion in instruction. The aim of this research is for providing with all the inclusive information by applying various machine learning models. For analysing the data, this paper has provided various methods: 1. Avocado type prediction by Random forest and Decision tree. 2. Bike Sharing- prediction the count of bike usage by Multiple-linear Regression. 3. Adult dataset- for predicting the class of people based on their salary by Multiple-Logistic Regression and Support vector machine.**

Keywords— machine learning models, classification algorithms, Multiple-linear model, Random forest, Decision Tree, Multiple-Logistic Regression, Support vector machine.

I. INTRODUCTION

A. Avocado Prediction

The most important feature of the avocado is its appearance. The appearance of the fruit influences the market as well as the choice of the customer and quality to a certain extent. For checking the quality of food, certain parameters are taken into a measure: size, shape, colour, and texture, etc. Manual testing for quality of avocado is very time consuming and it also increases labour work manually. Therefore, quality control for avocado and all the related food products are done with computer systems which are widely used in industries [18]. The machine learning model and computer vision technology used for quality checks are based on data analysis and image processing. The process and an overview are explained regarding classification are given in this paper. Common features of the avocados are explained with some methods as “Speeded Up Robust Features (SURF)”, “Histogram of oriented gradient (HOG)”, and “Local Binary Pattern (LBP)” like size, shape, texture. Random forest and Decision tree models are used. With the help of Random forest and Decision tree good performance was observed for the model.

B. Bike Sharing

Predicting the count of the bike can help the people as well as the company to decide when it is the best time to ride the bike and for the company, it is best to know the count of bikes so they can implement it which will be profitable. Generally, a bike-sharing system is used normally for day to day as public transport and a short distance journey [11]. Using a bike-sharing system makes sure that the pollution is reduced as compared to the use of motor vehicles that emit pollutants into the air [4]. It arranges for the number of users, start time, end time, number of users, travel journey, pick-up location, and drop-off location and the last thing price tag will specify the count of bike-sharing. The price can be measured with the distance and the count of bike used, but the price calculation is not a trivial task over here. The multiple-linear Regression (LR) model is used on this dataset for achieving the required results.

C. Adult dataset

In the United States, right now the most important concern is about the prominent inequality of income and wealth is a big thing. The world’s rushing economic inequality is reduced due to the likelihood of decreasing poverty. Sustainable improvement and development in economic stability depend on the universal principle of moral equality [14]. For addressing the problems and providing the solution, the government in different countries are trying their best to identify it and reduce it. The main aim of this study and the dataset is to show the usage of the machine learning models and data mining techniques for helping humans with a solution to the income problem. The adult dataset from UCI is used for the study purpose over here. The classification model is applied over here for analysing whether the people’s yearly income falls under 50k or above 50k category based on all the variables [6]. Logistic model and support vector machine are applied on this dataset; but the SVM helped with the highest accuracy of 84.7%.

II. RELATED WORK

Hereafter, the primary attention will be on ideas which use bike-sharing and are mainly related to our question of the count prediction problem.

The market price of tomato and the size of the tomato have a relationship with each other. If tomatoes are sorted manually which are dependent on human explanation are most likely to have an error. Python programming is used for classification of tomato as small in size, medium in size, and large by looking at a single tomato with the implementation of techniques such as machine learning and deep learning [1]. Area and radius and perimeter information are gathered of tomatoes by observing the images of different sizes. The output of the experiment explains that using thresholding, the accuracy of 85.83%, and 80% is attained for radius, perimeter and the area. By using machine learning, the accuracy of 94% - 95% is achieved by applying SVM, 97.5% - 93.5% is achieved by KNN, and 90.33 - 92.5 is achieved by ANN. By observing the values, we can explain that SVM is the most dependable model without overfitting [1]. By looking at the deep learning values regardless the algorithm, low performance it obtained like 82.31% - 78.21% - 55.97% training and testing accuracy for VGG16, 48.17% - 41.11% - 37.64% for InceptionV3, and 56.05% - 44.96% - 22.78% for models. The analysis of comparison done by machine learning technique and deep learning about tomato fruit size in classification for checking accuracy performance.

Earlier as we discussed the classification of tomatoes, fruit classification is also done by visual inspection, it bears from a problem of unpredictability in prediction by different people. A smart machine is needed for classifying the fruits automatically because the classification of fruits by human beings is more costly [2]. This study helped in practically real-time classification of fruits to reduce the labour cost of the organisation. This system uses colour pictures for processing the method for fruits and infrared technology is used for internal factors. For this study, an artificial neural network model is used so that it can classify fruits. 1900 fruits are trained and tested by implementing ANN machine learning model [2]. The accuracy rate observed by applying ANN model is 97.5 in the experiment.

It can be observed from paper [1] that the author analyzed type prediction by applying KNN, random forest, SVM, ANN, etc. on tomato type prediction. SVM was applied with and accuracy of 97.5%. Visual inspection is time-consuming as well as not cost-friendly, so ANN was applied with an accuracy of 97.5%.

As the market is expanding around the whole world into the future generation of bike-sharing technology without docking stations is recently evolving the long-established bike-sharing market. This study of bike-sharing is used for getting information about the usage of dockless bike service across Singapore [3]. We have observed and gathered all the data of the dockless bikes from a company in Singapore of approximately a week of more than 13 million records. We considered spatial autoregressive models for analysing the spatiotemporal patterns during a time of period of bike usage. The weather conditions, access to public transport, the environment was explored by the models depending on the dockless bike's usage [13]. The diminishing marginal impact is related to a large group of bikes with the usage of a bike to a great extent. With the high impact, easy to access public transport, more promotions on a free ride which has a great influence on the usage of a dockless bike [3]. They will have a negative impact on rainfall and high temperature on the usage of the bike. This study of a dockless bike also gave us some guidance to policy changes, about transportation, sustainability, new practices. Similarly, in this paper, we can apply Multi-linear regression which is also a type of regression model. It is effective in predicting the count of dockless bikes used because of its strong ability to show the linear relationship between the variables and it is also used to check the outliers in the variables.

The usage of the bike-sharing system in the world is increasing as a great medium of transport. Almost all the developing counties have started using the bike-sharing system [4]. Earlier in India, several entrepreneurs attempted to established this system but did not succeed at that time because they were not implementing data analytics correctly. In some cases when the traveller comes to get a bike, the bike stations were empty, or as compared to the people the bikes were less. Therefore, this system will help to establish a system correctly with all measured and will also help the travellers for planning their travels. The predictions can be made every hour so that we will be able to get the data for peak hours for entering into the data system. Likewise, in this paper, we can use the classification model like a random forest which will help us to get a count for predicting the count of bike used per hour [4]. It will be useful for showing high accuracy within the variables from which we can provide the count of bikes. If the number of trees is more then it will not allow overfitting of trees in this model.

From the paper [3], the author did a comparative analysis of Multiple-linear regression, random forest on bike-sharing Count prediction and accuracy were observed by applying the linear model with a strong relationship with high accuracy. We can also observe, accurate real-time bike-sharing prediction can help us with a better travelling strategy [4].

["Pornthep K"] represents a prediction of the salary system using students who have graduated as a model. The technique of data mining is used for the generation of a model for predicting the salary of individual students having the same quality for training data. Considering the process, an experiment was performed on comparing five mining techniques like KNN, Decision tree, SVM, Naïve Bayes, and Neural networks for finding suitable techniques for salary prediction. By observing the data, we can see around 13541 records with a cross-validation method. By looking at the results we can observe that KNN gave a good efficiency as compared to other models for prediction of salary [5]. For making a judgment on the usage, the observed results were based on the system which was effective in motivating the students for studying and also gave them the track for positive thinking towards their future which was conducted with 50 users on a survey. The final output was observed and the satisfactory result was achieved by implementing this system because it was easy to use and results of prediction were easy to understand without learning about the background knowledge.

The author applied KNN, decision tree, naïve Bayes and neural networks called a multi-layer model [21]. This paper focuses on predicting the class dependent on the salary given by the companies [6]. A functional relationship is achieved by using a four-layer model between input and output. A very satisfying result was achieved and predicted error was very less [9]. Mean absolute error was used for evaluating the model with predicted and true values.

III. DATA MINING METHODOLOGY

A disciplined system development the performance of the model and the process can be adjusted is known as Knowledge Discovery in Databases (KDD) method [15]. This process include data cleaning, modelling, transformation, integration, pattern evaluation, and knowledge presentation; it is also known as an interactive process.

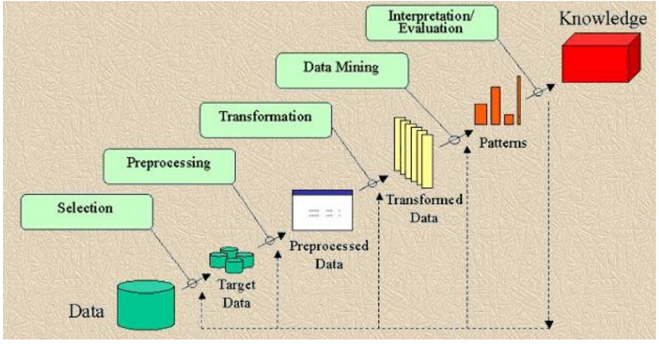


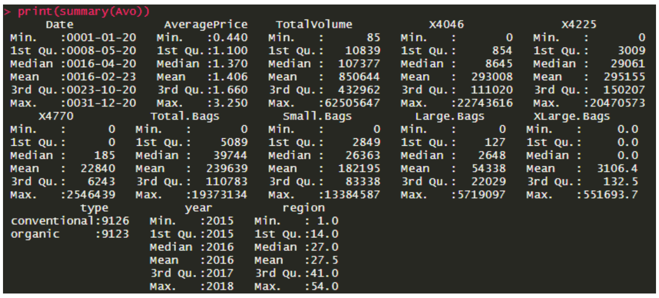
Fig. KDD outline process

A. Data Processing:

Usually the data that we collected have missing values, dummy variables, duplicate features, zero records. Therefore, for getting rid of such anomalies, data cleaning is done. From the below diagram we can see the steps of data processing to be done for getting accurate accuracy.

1. Avocado type prediction

This dataset is the real type prediction dataset of avocados. The table below will give a brief description of the data fields of this dataset. For predicting the accurate type of avocados, the anomalies should be removed and the data should be trained.



```
> print(summary(Avo))
      Date              AveragePrice    TotalVolume            X4046                X4225         
 Min.   :0001-01-20   Min.   :0.440   Min.   :      85   Min.   :       0   Min.   :       0  
 1st Qu.:0008-05-20   1st Qu.:1.100   1st Qu.:   10839   1st Qu.:     854   1st Qu.:    3009  
 Median :0016-04-20   Median :1.370   Median :  107377   Median :    8645   Median :   29061  
 Mean   :0016-02-23   Mean   :1.406   Mean   :  850644   Mean   :  293008   Mean   :  295155  
 3rd Qu.:0023-10-20   3rd Qu.:1.660   3rd Qu.:  432962   3rd Qu.:  111020   3rd Qu.:  150207  
 Max.   :0031-12-20   Max.   :3.250   Max.   :62505647   Max.   :22743616   Max.   :20470573  
     X4770             Total.Bags          Small.Bags           Large.Bags         XLarge.Bags      
 Min.   :      0   Min.   :       0   Min.   :       0   Min.   :      0   Min.   :     0.0  
 1st Qu.:      0   1st Qu.:    5089   1st Qu.:    2849   1st Qu.:    127   1st Qu.:     0.0  
 Median :    185   Median :   39744   Median :   26363   Median :   2648   Median :     0.0  
 Mean   :  22840   Mean   :  239639   Mean   :  182195   Mean   :  54338   Mean   :  3106.4  
 3rd Qu.:   6243   3rd Qu.:  110783   3rd Qu.:   83338   3rd Qu.:  22029   3rd Qu.:   132.5  
 Max.   :2546439   Max.   :19373134   Max.   :13384587   Max.   :5719097   Max.   :551693.7  
           type           year          region    
 conventional:9126   Min.   :2015   Min.   : 1.0  
 organic     :9123   1st Qu.:2015   1st Qu.:14.0  
                     Median :2016   Median :27.0  
                     Mean   :2016   Mean   :27.5  
                     3rd Qu.:2017   3rd Qu.:41.0  
                     Max.   :2018   Max.   :54.0  
```

Fig. Summary of Avocado dataset

From the above figure, we can see that: there are 13 variables within which some variables are anonymous and some are specified. The above tables will help you to observe the range of minimum, 1st quadrant, median, mean, 3rd quadrant, and maximum. The maximum and the minimum values are used for understanding the nature of the table [7]. No negative value is present in this numerical data type column. The above summary shows there are no null values present in this data. Outliers were cleared for getting the better performance of the model.

We have used variables like size, quantity, price, region variables for predicting the type of avocado. After further analysis, we can observe that some anonymous variables contributed significant information and thus were used for analysis.

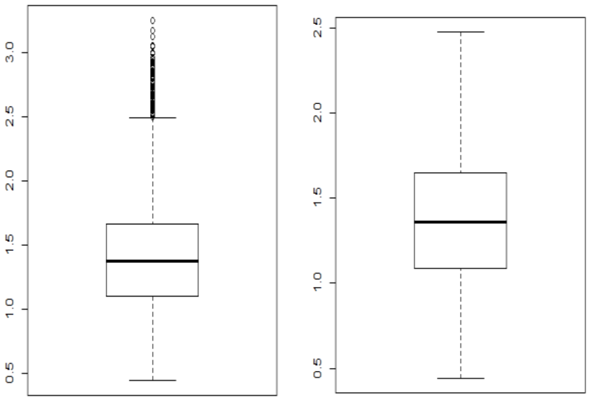


Fig. Identifying Outliers

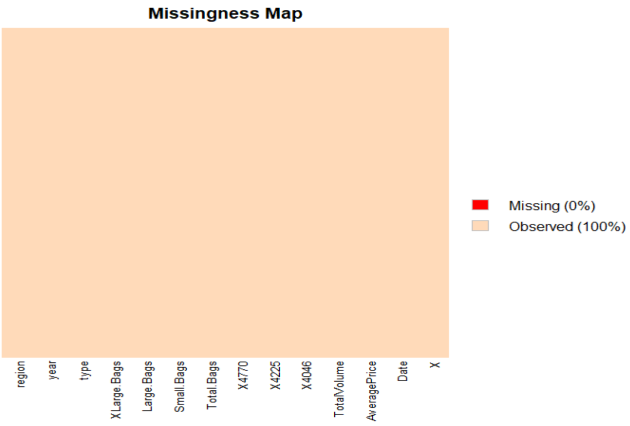


Fig. Heatmap with 0% Missing values

By looking at the above diagram, we can see that missing values were not present in the data. After this step we can conclude that the data is ready for performing the models.

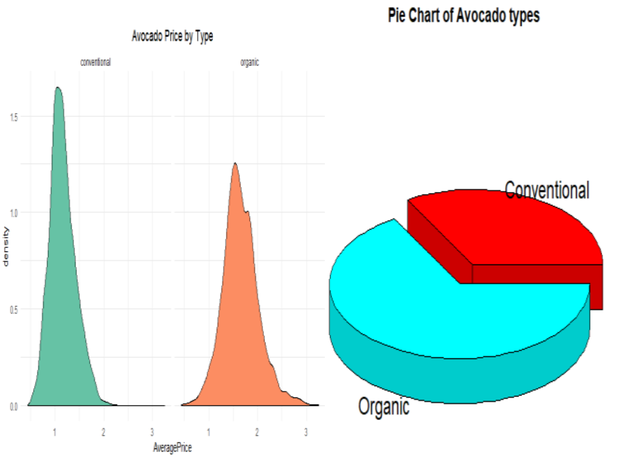


Fig. Type of density

From the above diagram on density and average price, we can see that conventional avocados were available at a greater quantity in the market than organic avocados. Average Price is one of the important factors used for predicting the type of avocado.

By looking at the plot below, we can see the price of the avocados is different for each type over 30days. The organic avocado is a little costly as compared to the conventional one. We can also notice the fluctuations in price.

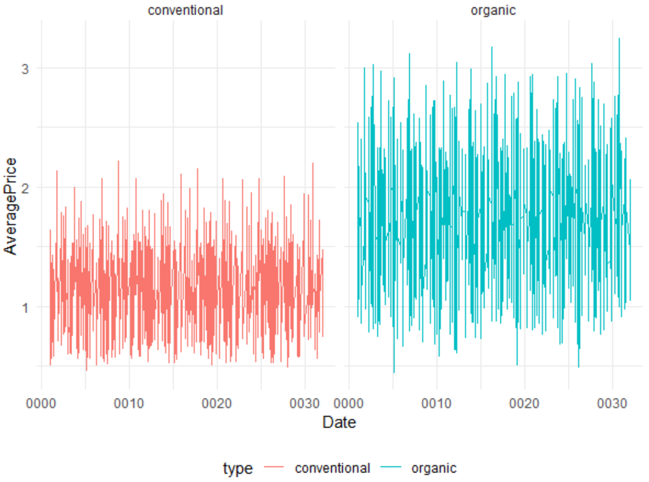
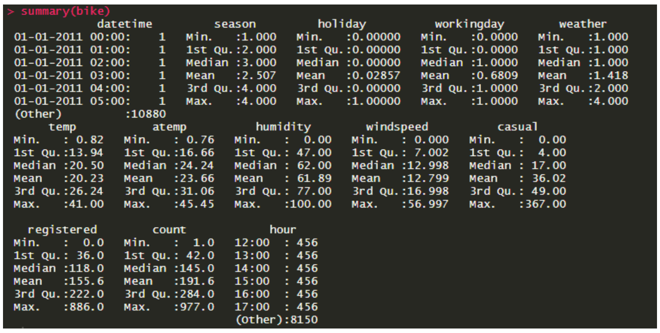


Fig. The price of each type

2. *Bike sharing*

The main goal of this data is to predict the count of bike sharing done with the other features. Working day and non-working days provide data on maximum use of bike-sharing systems with daily working hours of the day.

By looking at the summary we can see that, there are 13 variables present in the data from which we have to predict the count of bike sharing done. The nature of the column can be understood by observing minimum and maximum value. The correlation diagrams show the collinearity between the target variable and the remaining variable.



```
> summary(bike)
             datetime           season          holiday           workingday        weather     
 01-01-2011 00:00:    1   Min.   :1.000   Min.   :0.00000   Min.   :0.0000   Min.   :1.000  
 01-01-2011 01:00:    1   1st Qu.:2.000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:1.000  
 01-01-2011 02:00:    1   Median :3.000   Median :0.00000   Median :1.0000   Median :1.000  
 01-01-2011 03:00:    1   Mean   :2.507   Mean   :0.02857   Mean   :0.6809   Mean   :1.418  
 01-01-2011 04:00:    1   3rd Qu.:4.000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:2.000  
 01-01-2011 05:00:    1   Max.   :4.000   Max.   :1.00000   Max.   :1.0000   Max.   :4.000  
 (Other)         :10880                                                                       
      temp           atemp          humidity         windspeed          casual      
 Min.   : 0.82   Min.   : 0.76   Min.   :  0.00   Min.   : 0.000   Min.   :  0.00  
 1st Qu.:13.94   1st Qu.:16.66   1st Qu.: 47.00   1st Qu.: 7.002   1st Qu.:  4.00  
 Median :20.50   Median :24.24   Median : 62.00   Median :12.998   Median : 17.00  
 Mean   :20.23   Mean   :23.66   Mean   : 61.89   Mean   :12.799   Mean   : 36.02  
 3rd Qu.:26.24   3rd Qu.:31.06   3rd Qu.: 77.00   3rd Qu.:16.998   3rd Qu.: 49.00  
 Max.   :41.00   Max.   :45.45   Max.   :100.00   Max.   :56.997   Max.   :367.00  
                                                                                    
   registered        count              hour     
 Min.   :  0.0   Min.   :  1.0   12:00   : 456  
 1st Qu.: 36.0   1st Qu.: 42.0   13:00   : 456  
 Median :118.0   Median :145.0   14:00   : 456  
 Mean   :155.6   Mean   :191.6   15:00   : 456  
 3rd Qu.:222.0   3rd Qu.:284.0   16:00   : 456  
 Max.   :886.0   Max.   :977.0   17:00   : 456  
                                 (Other):8150  
```

Fig. Summary of Bike Sharing dataset

The training dataset is transformed for getting a balanced class distribution is known as data sampling. By this, we can an imbalanced class can also be addressed while data preparation. Data sampling was not applied here since there was no imbalanced class present.

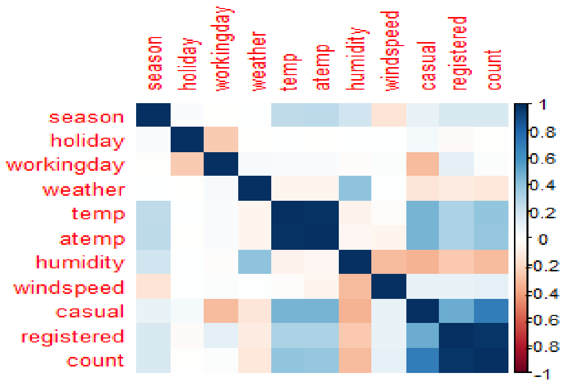


Fig. Correlation Diagram

From the heatmap diagram we can see that there were no missing values in this dataset. Like the earlier dataset, the data is organized properly over here, so the mapping process was not carried out. Hour Data has been separated from the date column for predicting count. Outliers were observed during the process which was then removed.

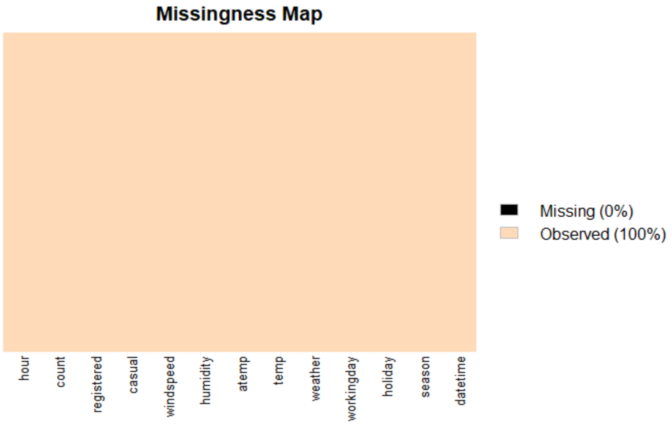


Fig. Heatmap with 0% Missing value

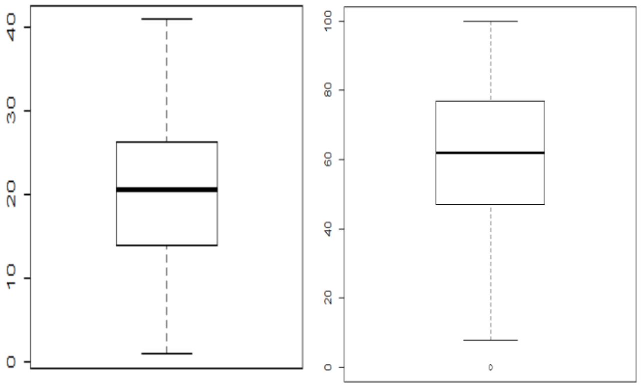


Fig. Outlier Identification

Featuring engineering was used on these plots for getting accurate insights into the data per hour. The most crucial step in predictive Modeling is feature engineering [18]. Figure 1 represents the usage of the bike on a working day. We can see that the usage of the bike in the morning around 9 am – 11 am is high and in the evening between 5 pm – 7 pm is high as compared to the rest of the day.

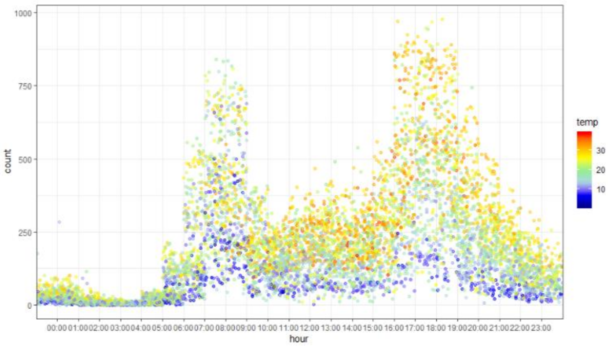


Fig. Insights into the usage of the bike during the working day/hour

In figure 2, the plot shows the usage of bike-sharing on a non-working day. From the plot, we can observe that the usage of the bike is high in the evening as compared to the rest of the day. The usage is least is the morning on a non-working day.

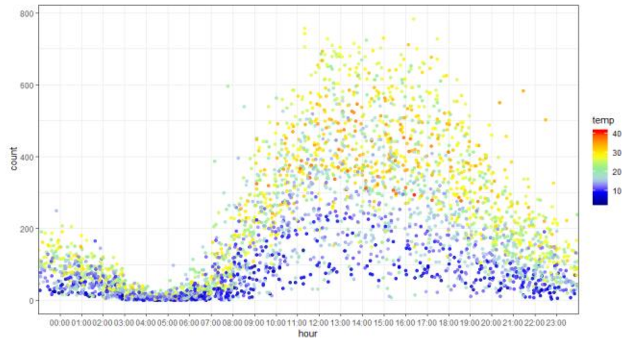
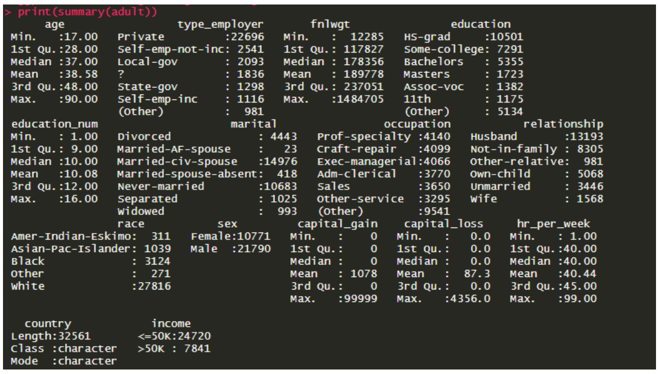


Fig. Insights of bike-sharing during non-working day

3. *Adult dataset*

This project aims to identify the class of people by comparing their salaries as well as work category, working hours per week, etc. Depending on the target variable, all the features were completely analysed upon their contribution was either omitted or kept.

Before working on this dataset, cleaning data is one of the important tasks with processing techniques. Missing values are present somewhere in the form of categorical, occupation, country, etc. which is taken care of with the help of transformation in algorithms of data.

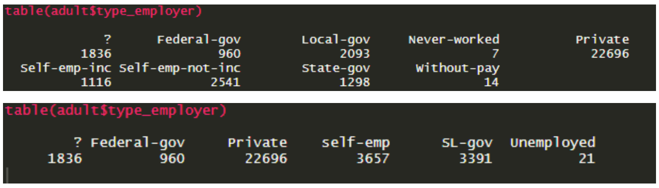


```
> print(summary(adult))
      age                type_employer        fnlwgt                education    
 Min.   :17.00   Private         :22696   Min.   :  12285   HS-grad     :10501  
 1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827   Some-college: 7291  
 Median :37.00   Local-gov       : 2093   Median : 178356   Bachelors   : 5355  
 Mean   :38.58   ?               : 1836   Mean   : 189778   Masters     : 1723  
 3rd Qu.:48.00   State-gov       : 1298   3rd Qu.: 237051   Assoc-voc   : 1382  
 Max.   :90.00   Self-emp-inc    : 1116   Max.   :1484705   11th        : 1175  
                 (Other)         :  981                     (Other)     : 5134  
 education_num                    marital                occupation            relationship  
 Min.   : 1.00   Divorced             : 4443   Prof-specialty :4140   Husband       :13193  
 1st Qu.: 9.00   Married-AF-spouse    :   23   Craft-repair   :4099   Not-in-family : 8305  
 Median :10.00   Married-civ-spouse   :14976   Exec-managerial:4066   Other-relative:  981  
 Mean   :10.08   Married-spouse-absent:  418   Adm-clerical   :3770   Own-child     : 5068  
 3rd Qu.:12.00   Never-married        :10683   Sales          :3650   Unmarried     : 3446  
 Max.   :16.00   Separated            : 1025   Other-service  :3295   Wife          : 1568  
                 Widowed              :  993   (Other)        :9541                          
                 race            sex          capital_gain    capital_loss     hr_per_week  
 Amer-Indian-Eskimo:  311   Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00  
 Asian-Pac-Islander: 1039   Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00  
 Black             : 3124                  Median :    0   Median :   0.0   Median :40.00  
 Other             :  271                  Mean   : 1078   Mean   :  87.3   Mean   :40.44  
 White             :27816                  3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00  
                                           Max.   :99999   Max.   :4356.0   Max.   :99.00  

   country             income     
 Length:32561       <=50K:24720  
 Class :character   >50K : 7841  
 Mode  :character                
```

Fig. Summary of the Adult dataset

The summary table helps us with the information of all the columns with predictable columns. For accessing object variables, R has a very powerful indexing feature built-in it. Selecting and excluding elements of observation can be done by using this feature. The performed code shows to keep or remove the variables and observations for getting random samples.



```
table(adult$type_employer)

               ?      Federal-gov        Local-gov     Never-worked          Private 
            1836              960             2093                7            22696 
    Self-emp-inc Self-emp-not-inc        State-gov      Without-pay 
            1116             2541             1298               14 
```

```
table(adult$type_employer)

         ? Federal-gov     Private    self-emp      SL-gov  Unemployed 
      1836         960       22696        3657        3391          21 
```

Fig. Subsetting of data

Subsetting was implemented in this dataset. In this some variables were removed for increasing the performance of the dataset and some were clubbed. Subsetting is used for speeding up the process. Subsetting of countries have been performed here by clubbing country into continents. Similarly, the subsetting of martial and employment status is also done. After modification of the data the missing data was reduced to one percent of the dataset and later on it was cleaned.

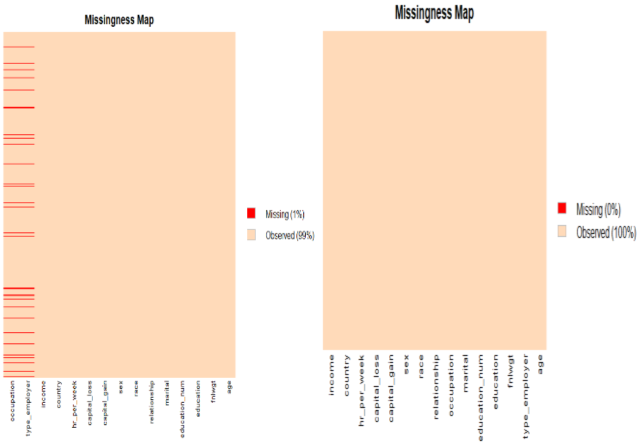


Fig. Heatmap with 1% and 0% Missing values

From the figure mentioned below, we can get an insight into a particular country with an income count. We can observe that North America has more people who are earning more as well as earning less than 50K compared to other continents.

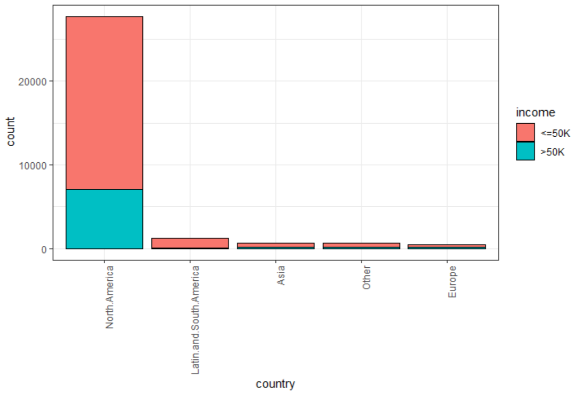


Fig. Bar plot of a country with an income

The histogram is generally used for displaying the distribution of variables, for comparison variables bar charts are used. In this graph, we are displaying the income of the people depending on their age. People having age between 25-50, are people with more income as compared to other aged people. People above 75years have the least income.

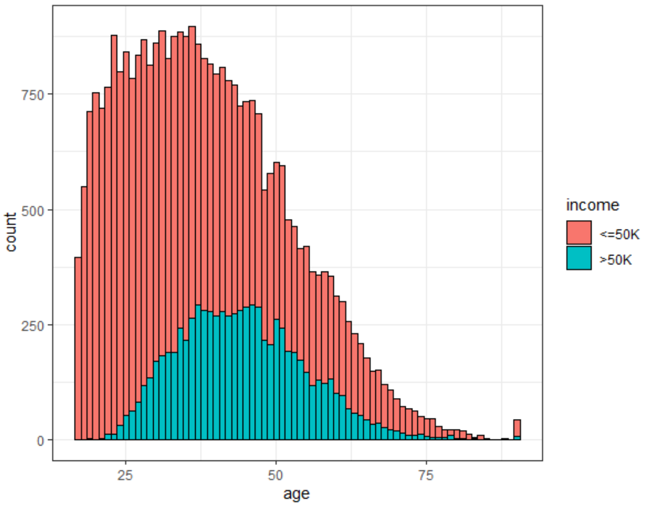


Fig. Histogram of age and income

B. Working Methodology:

The working section explains the model built on data that has inputs and the required output with a machine learning approach. Machine learning has instances like regression and classification.

1. Avocado dataset:

The dependent variable to be predicted is 'type' which consists of organic and conventional type. A classification model will be a perfect choice for this project as the variable which we are going to predict is categorical. With the help of observed research, two following classification machine learning models will be applied to this data as:

Decision Tree:

Decision trees are used for classification and regression, which are supervised non-parametric learning methods. If and else decision rule is used to set a sine curve from data by applying Decision trees. The fitter the model, the deeper the tree and more complex the decision rules. CART algorithm is used while performing a decision tree [12]. In this model, the dataset breaks down into many subsets and a decision tree is developed at the same time [10]. The final result is decision nodes and leaf nodes built in a tree. This model can look after numerical as well as categorical data.

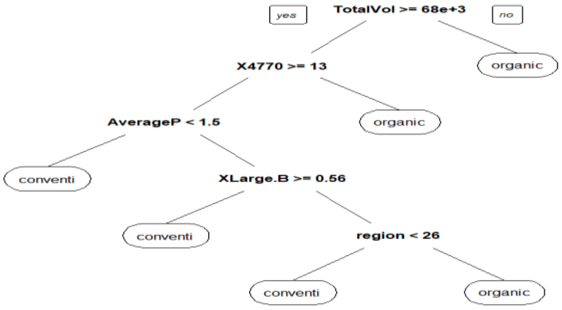


Fig. Decision tree of Avocado dataset

The outcomes of applying the decision trees:

- Decision trees will outperform as compared to the regression model if we can see high non-linearity and complex relation within dependent and independent variables [22].
- If you want to explain a model to people easily, the decision tree will always perform better than a linear model.

Random Forest:

The second model applied to this data set is random forest. The group of machine learning methods used for classification and regressions of constructing a large number of decision trees. It inspects at every split in through the process, a random subset of a feature [16]. Correlation between the trees is avoided by this process [21]. If our predictor variable is strong then this model will use these values during the first split. All the trees will be similarly displayed. Therefore, all the predictions will be highly correlated. Correlated predictors are not used for improving accuracy because of this the random forest systematically avoids correlation and increases performance. This model also performs with more accuracy than decision tree as aggregated result is produced which seems to be more accurate.

2. Bike Sharing:

Here, various elements like wind speed, temperature, humidity, the weather will be observed and we will predict the count of bikes used in all these elements. Multiple-linear Regression is applied to this dataset.

Multi-Linear Regression:

The multi-linear machine learning model is one of the easily understood models but still a very powerful algorithm used for modelling a strong linear relationship within the dependent and independent variables. For predicting the outcome of the different variables, linear regressions use many variables. This model looks after the multiple independent variables related to the dependent variable. In this dataset, this model is used for predicting a continuous count of bikes considering the remaining variables.

3. Adult Dataset

Here, the classification model is preferred on this dataset because the target variable is categorical. Two machine learning models are applied to this dataset as follows:

Logistic Regression:

Logistic regression is a machine learning model that is applied to this dataset which targeted the categorical variables. It is also used to measure the parameters of a logistic model which is a form of binary regression. A continuous value is observed which is the probability of the model and later on it is converted into target class [19]. Mathematically, dependent variables are present in a binary model such as 'pass and fail' which are called as indicator variables, whereas the values are labelled as '0' and '1'. The main reason for applying this model is that its estimated value is well measured. In short, this model is a logistic function of the linear combination. Therefore, we can determine that this model is very fast and highly accurate for understanding the way of prediction. Logistics can be binomial, ordinal, or multinomial.

Support Vector Machine:

This is the second model applied to this dataset. Classification of the algorithm and two group classification problems can be done by implementing supervied model known as Support vector machine. It works on the principle of a maximum "marginal classifier" [17]. "N-dimensional space" which classifies the data points because SVM generates a linear and non-linear hyperplane in an N-dimensional space. SVM also have a feature of being robust towards outlier that could also generate a hyperplane with the highest margin implementing this model i.e. nearest data points. Hence, we can conclude that the SVM model and logistic model give almost similar accuracy to this dataset.

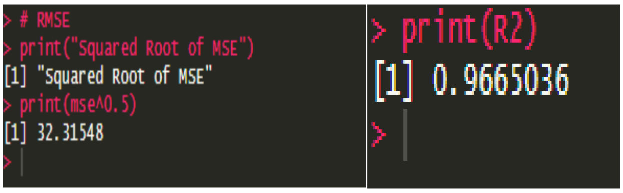
V. EVALUATION AND RESULT

A. *Root Mean Squared Error:*

Root mean squared error (RMSE) is used in regression models. The magnitude of the RMSE is measured by rule names as a scoring rule. Variation is observed after applying RMSE between Actual values and Predicted values [20]. Squaring the difference between these is necessary because sometimes we observe dissimilarity which can be negative or positive, to remove these dissimilarities we square the values. The root is applied to match the order of data with the order of loss function.

- *Bike sharing:*

The RMSE diagram shows that Root Mean Square of around 32.31 was obtained for the count prediction of bike sharing. Observed R-squared value is 0.966. We can notice that the dependent and independent variables are highly correlated.



```
> # RMSE
> print("Squared Root of MSE")
[1] "Squared Root of MSE"
> print(mse^0.5)
[1] 32.31548
> 
```

```
> print(R2)
[1] 0.9665036
> 
```

Fig. RMSE value and R-square value

B. *Confusion Matrix:*

For determining the accuracy and the error rate for predicting the algorithm, a confusion matrix is used. It is applicable for almost all the predictive analytics models (regression and classification alike). This accuracy is calculated using the trained model on the test data set.

1. *Specificity and Sensitivity:*
   - Sensitivity: The proportion of positive data that is correctly classified by the sensitivity model. The accuracy is of the model is explained as a factor.
     TP/(TP+FN)
   - Specificity: We can see the ratio of negative data values with total number of negatives and positive values.
     TN/(TN+FP)
2. *Precision and Recall:*
   - Precision: The prediction of the data in a correctly classified format can be specified as the probability of a model due to precision.
     TP/(TP+FP)
   - Recall: The completeness of the estimated result is observed by applying a recall of model which can be considered as same as sensitivity.
     TP/(TP+FN)
3. *Mis-classification:*
   The miss-classification error classifies the variable which is below a category into a different category. The best method for minimizing the error is to reduce it.
   (FN+FP)/(TP+TN+FN+FP)

Models	Accuracy	Specificity	Precision	Miss-Classification	Recall/ Sensitivity
Decision Tree	0.9594	0.9477	0.9489	0.040	0.9711
Random Forest	0.9964	0.9950	0.9978	0.0035	0.9950
SVM	0.8479	0.5803	0.8706	0.152	0.9367
Logistic Regression	0.8459	0.6200	0.9293	0.154	0.9728

Fig. Performance of Models

Avocado dataset:

As we see, Decision tree and Random forest model have been implemented on this dataset. By looking at the values we can say that we received a better performance by using random forest as compared to decision tree. (accuracy of Random Forest = 99.64%)

Adult dataset:

As we see, Logistic Regression and SVM model is applied to this dataset. By looking at the values we can say that we received a better performance by using SVM model as compared to Logistic Model. (accuracy of SVM = 84.79%)

VI. CONCLUSION AND FUTURE WORK

In this research we have implemented five different models that were applied on three datasets. In "Avocado Prediction" dataset, a 5% higher result was achieved with Random forest than Decision tree. We can predict the type of avocados by machine more quickly and efficiently as compared with human labour. In future these techniques should be used for reducing the processing time and increasing work efficiency. In case of "Bike Sharing" dataset, with a large repository, we observed a good result by getting the count of bikes with other factors from the dataset. In future we could increase the count of bikes in a peak hour and by looking at the weather. From "Adult" dataset, Logistic model and SVM provided a great classification result within the people class. In future more factors can be used with different models for prediction considering more countries.

The empirical searching from this research can give us opportunities for future work in implementing the classification models by normalization the data, tuning parameters, collecting the methods. Most importantly, using an advanced way such as neural networks for classification and incorporating hybrid approaches.

VII. REFERENCES

[1] Robert G., Elmer P. "Size Classification of Tomato Fruit Using Thresholding, Machine Learning, and Deep Learning Techniques", *Journal of Agricultural Science*, DOI:.10.17503/agrivita. v41i3.2435. Corpus ID: 208119310, 2019.

[2] H. Choi, J. Cho "A Real-time Smart Fruit Quality Grading System Classifying by External Appearance and Internal Flavor factors", *IEEE International Conference on Industrial Technology (ICIT)*, DOI: 10.1109/icit.2018.8352510. Corpus ID: 13747028, 2018.

[3] M. Gan, Daben Y. "A deep learning approach on short-term spatiotemporal distribution forecasting of dockless bike-sharing system", *Neural Computing and applications*, DOI: 10.1007/s00521-018-3470-9, 2018.

[4] A. Patil, K. Musale "Bike Share Demand Prediction using RandomForests", *International Journal of Innovative Science, Engineering & Technology*, Vol. 2 Issue 4. ISSN 2348 – 7968, 2015.

[5] Pornthep K, Pokpong S. "Implement of salary prediction system to improve student motivation using data mining technique", *11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, DOI**:** 10.1109/KICSS.2016.7951419, 2016.

[6] I. Martín, Andrea M. "Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study", *International Journal of Computational Intelligence Systems*, Volume 11, Issue 1, 2018.

[7] F. Liu, L. Snetkov, "Summary on fruit identification methods: A literature review", *Advances in Social Science, Education and Humanities Research*, volume 119, 2017.

[8] H. Singh,"Understanding Random Forest", *One Algorithm at a time,* 2019.

[9] A. Ghanem, M. Elhenawy "Modeling Bike Availability in a Bike-Sharing System Machine Leaning", *Conference Paper*, 2017.

[10] N.S. Chauhan, "Decision Tree Algorithm – Explained", *Towards Data Science,* 2019.

[11] O.P Nekkanti "Prediction of Rental Demand for a bike share program", *North Dakota State University of Agriculture and Applied Science*, 2017.

[12] C. Sehra, "Decision Tree Explained Easily", *Medium Publication,* 2018.

[13] S. Kumar, Yongyun C. "A rule-based model for Seoul Bike Sharing demand prediction using weather data", *Deep learning for Remote Sensing Environments*,2020.

[14] Navoneel C, S. Biswas "A Statistical approach to adult census income level prediction", *Jalpaiguri Government Engineering College*, 2014.

[15] A.L Buczak, E. Guven " A Survey of data mining and machine learning methods for cybersecurity intrusion detection", *IEEE Communications Surveys & Tutorials,* DOI: 10.1109/COMST.2015.2494502, Volume 18, Issue-2, 2016.

[16] Synced. "How Random Forest Algorithm works in Machine leaning", *Medium Publication*, 2017.

[17] Susan Li "A comprehensive survey on support vector machine in data mining tasks: Applications and challenges", *International journal of database theory and application,* Vol 8, No.1 (2015), pp.169-186, 2015.

[18] S. Naik, B. Patel "Machine Vision-based Fruit Classification and Grading – A Review", *International Journal of Computer Applications,* DOI: 10.5120/Ijca2 017914937. Corpus ID:40750417, 2017.

[19] J.C Houwelingen, S. le Cessie "Logistic Regression, a review", *Wiley Online Library,* pp. 215-232, 1988.

[20] T. Chai, R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature", *An interactive open-access journal of the European Geosciences Union,* vol. 7*,* Issue. 3, pp. 1247-1250, 2014.

[21] E. Ngai, X. Sun "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature", *Decision Support Systems*, Volume 50, Issue 3, 2011.